社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2003－ＳＬＰ－49 (15)
2003／12／18

バイリンガル話者音声に基づく二言語混合音響モデルの話者適応法の検討

小笠原洋一† 大河雄一‡ 鈴木基之† 伊藤彰則† 牧野正三†

†東北大学大学院 工学研究科 〒980-9575 仙台市青葉区荒巻字青葉05
‡東北大学大学院 教育情報学研究部・教育部 〒980-8576 仙台市青葉区川内

E-mail: † {ogasawara, moto, aito, makino}@makino.ecei.tohoku.ac.jp, ‡ kuri@ei.tohoku.ac.jp

あらまし 本稿では，外国語発音学習システムに用いる二言語混合音響モデルの話者適応法について検討する．学習者は外国語を発音する際，母国語の適当な音素に置換して発音する傾向にあることから，認識の際には学習対象と母国語の二言語混合の音響モデルを用いる．音響モデルの精度向上のために話者適応が行われるが，通常，ある言語の音響モデルの話者適応を行うにはその言語の発声データを用いる必要がある．しかし，外国語発音学習システムにおいては学習する言語の正しい発声を得ることができない．そこで，本研究では学習者の母国語の発音データを用いた二言語混合音響モデルの話者適応方法を提案した．本報告ではさらに認識率の向上を目指し，二言語のバイリンガル話者の発声データを利用した話者適応方法を検討し，その評価を行った．評価実験の結果，従来の話者適応方法よりも高い性能が得られた．

キーワード 非母国語音声認識，二言語音響モデル，話者適応，MLLR，話者適応学習法

Speaker Adaptation of Bilingual Phone Models using Bilingual Speakers' Speech

Hirokazu OGASAWARA† Yuichi OHKAWA‡ Motoyuki SUZUKI†
Akinori ITO† and Shozo MAKINO†

† Graduate School of Engineering, Tohoku University 05 Aoba, Aramaki, Aoba-ku, Sendai, 980-9579 Japan
‡ Research and Education Divisions, Graduate School of Educational Informatics,Tohoku University
Kawauchi, Aoba-ku, Sendai, 980-8576 Japan

E-mail: † {ogasawara, moto, aito, makino}@makino.ecei.tohoku.ac.jp, ‡ kuri@ei.tohoku.ac.jp

Abstract In this paper, we investigate a method of speaker adaptation of bilingual phone models to improve precision of non-native speech recognition system. Non-native speakers tend to substitute native-language's phones for non-native phones, therefore the recognition system must use bilingual phone models consist of all phones in non-native and native languages. Speaker adaptation, generally, use utterance of the same language as the phone model. However, non-native speaker can't speak well to use speaker adaptation. In order to adapt bilingual phone models, we propose a speaker adaptation method of bilingual phone models using native speaker's utterance. To improve bilingual phone models, we propose a method using bilingual speakers' speech. Experiments showed that the bilingual phone models adapted by the proposed method outperformed the models adapted by conventional methods.

Keyword Non-native speech recognition, Bilingual phone models, Speaker adaptation, MLLR, Speaker-Adaptedve Traning

1. はじめに

国際化社会とよばれる現在，外国人と交流する機会が増えている．その際，コミュニケーションの手段として英語を用いることが多い．近年では外国人を指導教官とした英会話学習が盛んに行われている．指導者に直接発音を聞いてもらい，その場で発音誤りが指摘されるこの指導方法は理想的な学習方法であるが，時間的・金銭的制約が生じることは必至である．そこで，音声情報処理技術を活用して，学習者の発音の問題点を自動的に抽出し，それに従ってどのように発音を矯

正するかを具体的に学習者に示す，インテリジェントでかつインタラクティブな計算機援用発音教育システム（CALL システム）の開発が進められている．

このような外国語発音学習システムにおいては，発音を評価する際，学習者の発声した音声のどの部分がどの音素に対応しているかを学習者に示し，その部分を学習者がどのように発声したか，正しい発声をするにはどのように直せばよいのかを教えることが重要である．本研究では，学習者の音声入力に対して発音の誤り個所を指摘し，その誤りの対処法を指示する非母国語話者用音素ラベリングシステムをベースとした発音評価システムの構築を目的としている．

学習者がある言語の発声をする際の発音誤りの傾向として，学習する言語の音素を学習者の母国語の適当な音素に置換する[1]ことから，CALL システムでは，学習対象の言語の音響モデルと，学習者の母国語の音響モデルの，２つの異なる音素体系が混合している音響モデルを利用して音素ラベリングを行う．

音素ラベリングの精度は音響モデルの精度に依存するため，音響モデルの高精度化のために話者適応が用いられる．通常，ある言語の音響モデルの話者適応を行うにはその言語の発声を用いる必要がある．しかし，CALL システムの場合は目的が発音学習であるため，学習する言語の正しい発声を得ることができない．実際に学習者の不完全な発声を用いて適応を行うと，認識精度が悪化する[2]．

そこで本研究では，正しい発声が可能な学習者の母国語の発声を用いて，学習対象言語の音響モデルを話者適応する方法の検討を行った．本研究では日本人学習者が英語発声を学習するシステムを想定している．その際に用いる日本語と英語が混合している二言語混合音響モデルの話者適応について検討を行った．

2. 日本語英語混合音響モデルの話者適応

日本語と英語の混合音響モデルの話者適応について我々は，移動ベクトル場平滑化（VFS）法[3]を用いた話者適応方法を提案した．日本語英語バイリンガル話者の日本語発声で英語音響モデルの話者適応を行い，その有効性を示した[4]．そこで，この方法を用いて日本人話者の日本語発声による日本語英語混合音響モデルの話者適応を行った．また，話者適応に広く用いられる MLLR 法[5,6]を用いて，同様に日本語発声による日本語英語混合音響モデルの適応を行い，２つの方法の比較を行った．なお本研究における話者適応は，音響モデルとして隠れマルコフモデル（HMM）を用い，その平均ベクトルのみを更新を行うこととする．

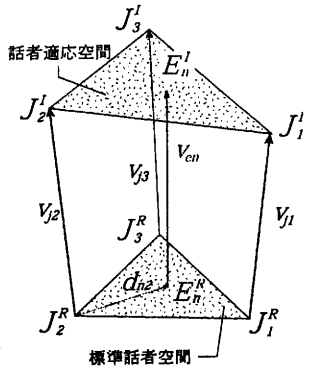

図１：VFS 法による英語音素移動ベクトルの推定

2.1. VFS 法を用いた適応

VFS 法は，音響モデルの学習の際，学習用データにない音素の移動ベクトルを，学習データから更新された近隣の音素の移動ベクトルから推定する方法である．この推定により，学習モデルに出現しない音素の適応を行うことができる．

適応用データを日本人学習者の日本語，適応する音素を日本語と英語とすると，VFS 法より近隣の日本語音素の移動ベクトルを用いて英語音素の移動ベクトル求めることができる．この様子を図１に示す．

2.2. MLLR 法を用いた適応

MLLR 法は，適応データから作成された変換行列を用いて線形変換を行うことにより，話者適応をする手法である．MLLR 法を用いると，適応データに出現しなかった音響モデル中の音素も適応することができる．

そこで，日本人学習者の日本語発声による日本語英語混合音響モデルの話者適応を行った．具体的な適応方法について以下に示す．

英語基準話者モデルの英語音素セット E，日本語基準話者モデルの日本語音素セット J があり，それぞれの音素セットに含まれるある音素の平均ベクトルを μ_E，μ_J とする．話者 N についての音素セット J の MLLR 回帰行列（線形変換行列 $A_{N,J}$ と平行移動係数 $b_{N,J}$ からなる行列）$W_{N,J}$ は次式のように求められる．

$$W_{N,J} = \left\{ \sum_t \gamma_t \cdot {\mu_J}^T \cdot {\Sigma_J}^{-1} \cdot \mu \right\}^{-1} \times \left\{ \sum_t \gamma_t \cdot {\mu_J}^T \cdot {\Sigma_J}^{-1} \cdot o_t \right\}$$

$$W_{N,J} = \left[A_{N,J} \quad b_{N,J} \right]$$

γ_t は HMM における入力学習データ $o_t^{(r)}$ の出力確率，Σ_J は種モデルの対角共分散行列を表す．

この回帰行列を用いて，音素セット E，J の平均ベクトルの更新を行う．各音素の平均ベクトルは次のように求められる．

$$\begin{cases} \mu_J' = A_{N,J} \cdot \mu_J + b_{N,J} \\ \mu_E' = A_{N,E} \cdot \mu_E + b_{N,E} \end{cases}$$

2.3. 性能評価実験

適応後の音響モデルの評価として，HTK Tool[7] の HVite を用いて，日本語英語混合音響モデルでの音素認識を行った．認識には言語モデルを用いていない．音声データは標本化周波数 16kHz，フレーム長 25ms，フレーム周期 5ms で分析を行い，用いた特徴量パラメータは 12 次元 MFCC，対数パワー，12 次元 ΔMFCC，Δ対数パワーの計 26 次元である．音響モデル，適応，認識用データの条件を表 2 に記す．音素認識における正解の音素系列は，母語話者 3 名による評価データの発音判定結果より作成した．この正解系列と音素認識結果より音素正解率を求め，適応による混合音響モデルの性能の比較を行った．

実験結果を図 3 に示す．VFS 法で適応すると，音素正解率は大きく低下してしまった．MLLR 法で適応をすると，適応前とほぼ同程度の音素正解率となった．この結果より，本研究においては MLLR 法について検討を行うことにする．

音素認識率が低下した原因として，適応に用いる基準話者モデル作成に用いた学習データが，日本語と英語で異なることが考えられる．本研究で用いた適応用基準話者モデルは，日本語不特定話者モデルは日本人の発声から，英語不特定話者モデルはアメリカ人の発

表 2 ：音響モデルの条件

音素HMM	monophone，4 状態 3 ループ 単一ガウス分布 日本語 28 音素，英語 52 音素 性別依存モデル
学習データ	日本語：ATR C セット 男性 32 名×100 文，女性 33 名×100 文 英語：TIMIT 男性 326 名×10 文，女性 136 名×10 文
適応用データ	日本人男性 2 名，女性 2 名 日本語発声　各 10～50 文
評価データ	上記の日本人話者　英語発声　各 10 文

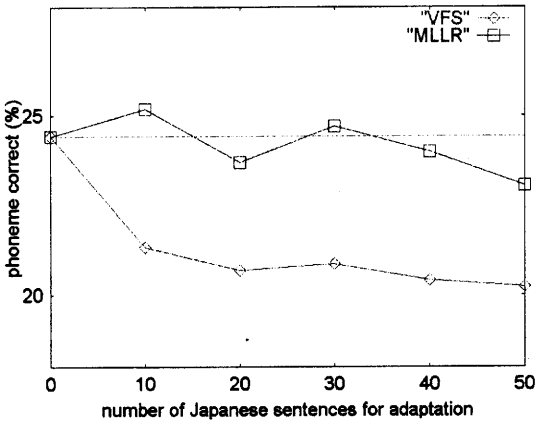

図 3 ：音素認識結果（VFS 法と MLLR 法の比較）

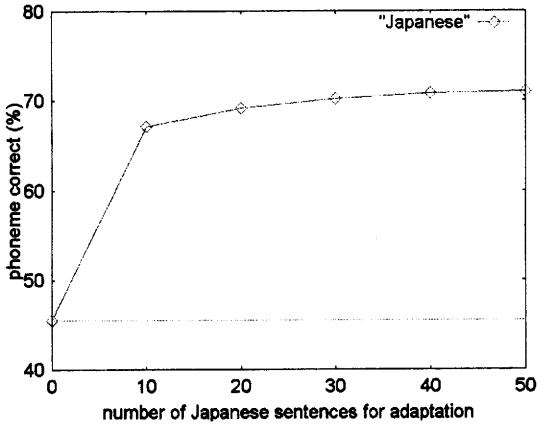

図 4 ：音素認識結果（適応後の日本語音響モデル）

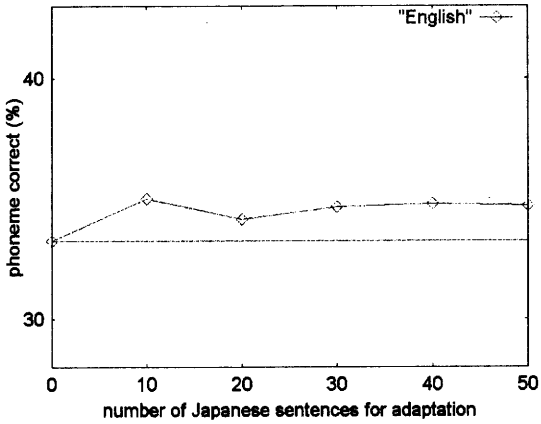

図 5 ：音素認識結果（適応後の英語音響モデル）

声から作成している．よって，日本語基準話者モデルの話者空間と英語基準話者モデルの話者空間が異なる．異なる話者空間に対して一方の言語の MLLR 回帰行列のみを用いて適応を行ったため，混合モデル全体としての精度が悪化してしまったと考えられる．

そこで，MLLR 法で適応を行った混合音響モデルを日本語と英語に分けて同様に音素認識を行い，言語による適応効果の差を調査した．その結果を図４，５に示す．日本語の音素認識率は20%以上の改善が見られたのに対し，英語の音素認識率の上昇はわずか2～3%程度であった．これより，日本語発声から作成された回帰行列を用いて適応を行うことで，英語音素の適応の効果が得られることがわかった．しかし，日本語音素と英語音素の適応効果に大きな差が生じたために，混合音響モデルのバランスが崩れてしまい，全体として認識率が下がったと考えられる．よって，混合音響モデルを MLLR 法で適応する際には，日本語音素と英語音素の適応に用いる回帰行列はそれぞれ別のものを用いる必要がある．

3. バイリンガル話者空間の写像を用いた適応

話者適応に用いる不特定話者モデルの話者空間が言語によって異なることから，日本語音素と英語音素それぞれの適応には異なるMLLR 回帰行列を用いる必要がある．学習者から正しい英語発声を得ることができないため，英語音素適応のための回帰行列を推定しなければならない．そこで，両言語を正確に発声することができるバイリンガル話者空間を経由することにより，日本人学習者の英語音素適応のための回帰行列を推定する方法を提案する[8]．

3.1. 適応方法

２つの音素セット E，J があり，それぞれの音素セットに含まれる適当な音素の平均ベクトルを μ_E，μ_J とする．話者 B に対する音素セット E，J の MLLR 回帰行列の線形変換行列をそれぞれ，$A_{B,E}$，$A_{B,J}$，平行移動係数を $b_{B,E}$，$b_{B,J}$ とする．

ここで音素セット E の話者空間から音素セット J の話者空間へ，線形写像によって変換できると仮定する．このとき，音素セット E から音素セット J への写像を $T_{E\to J}$，音素セット J から音素セット E の写像を $T_{J\to E}$ とすると，次のように表すことができる．

$$\left(A_{B,J},b_{B,J}\right)=\left(A_{B,E},b_{B,E}\right)\cdot T_{E\to J}$$
$$\left(A_{B,E},b_{B,E}\right)=\left(A_{B,J},b_{B,J}\right)\cdot T_{J\to E}$$

ここで，$T_{E\to J}=T_{J\to E}^{-1}$ である．

新たな話者 N について同様に考える時，音素セット J の線形変換行列 $A_{N,J}$ と平行移動係数 $b_{N,J}$ だけが分かっているとする．このとき，${}_NA_E$ と ${}_Nb_E$ について，写像を用いて以下のように表すことができる．

$$\begin{aligned}\left(A_{N,E},b_{N,E}\right)&=\left(A_{N,J},b_{N,J}\right)\cdot T_{J\to E}\\&=\left(A_{N,J},b_{N,J}\right)\cdot\left(A_{B,J},b_{B,J}\right)^{-1}\cdot\left(A_{B,E},b_{B,E}\right)\end{aligned}$$

すなわち，いったん音素セット E の平均ベクトルを話者空間 B に写像し，そこから音素セット J に戻した上で ${}_NA_J$ を用いて話者空間 N に写像することになる．

この結果，話者 N について更新された音素セット E の平均ベクトルは，

$$\begin{aligned}\mu_{N,E}'&=A_{N,E}\cdot\mu_E+b_{N,E}\\&=\left[A_{N,J}\cdot A_{B,J}^{-1}\cdot A_{B,E}\right]\cdot\mu_E\\&\quad+\left[A_{N,J}\cdot A_{B,J}^{-1}\cdot\left(b_{B,E}-b_{B,J}\right)+b_{N,J}\right]\end{aligned}$$

のように求まる．

3.2. 性能評価実験

音響モデルの評価として，2.3.節と同様の音素認識実験を行った．適応話者の英語音響モデルに対する MLLR 回帰行列の推定には，日本語英語バイリンガル話者３名(男性１名，女性２名)の日本語 10～50 文章，英語 10～50 文章を用いた．各個人の発声から回帰行列を作成したもの（M01, F01, F02），女性２名の発声から回帰行列を作成したもの（female），全員の発声から回帰行列を作成したもの（all），計５種類の話者空間の写像を経由し，その音素正解率を求めた．それ以外の条件は，すべて 2.3.節に提示した条件と同じである．

各バイリンガル話者について各言語 30 文章から作成した回帰行列を用いて話者適応をした音響モデルの音素正解率を図６に示す．baseline は 2.3.節における MLLR 法による適応の音素正解率である．

バイリンガル話者空間の写像を用いて適応すると，適応前と比較して音素正解率は最大で５％程度上昇した．経由する話者空間によって認識精度が異なることから，この方法で適応をする際には，適切な話者空間を選択する必要がある．

また，female は F01, F02 よりも音素正解率が低く，all は音素正解率が最も悪かった．いずれも複数人の発声データから回帰行列を作成したものであることから，発話データのばらつきが音素正解率に影響していると考えられる．

4. 話者適応学習法を用いた適応

写像作成に用いるバイリンガル話者によって適応

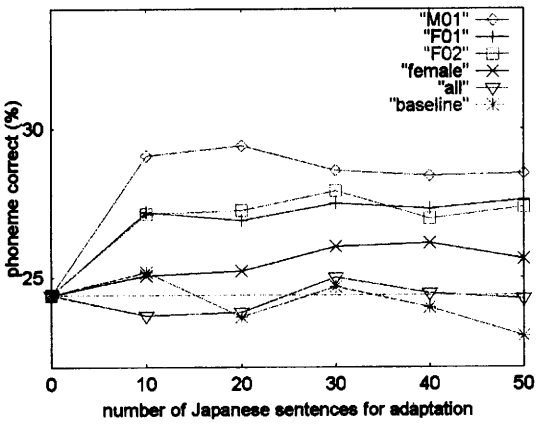

図6：音素認識結果（バイリンガル話者空間を経由）

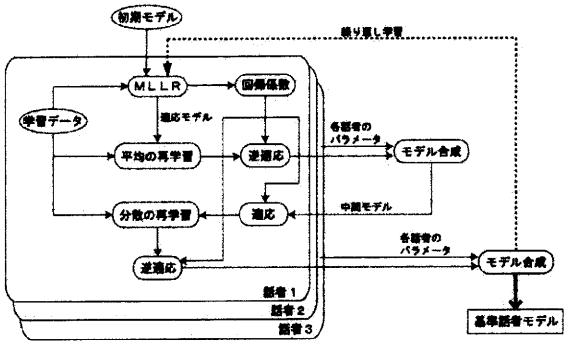

図7：SAT法の学習手順

の効果が異なることから，共通したバイリンガル話者空間，つまり，バイリンガル話者間の話者性を取り除いたバイリンガル話者全体を表す話者空間の作成が必要になる．そこで，話者適応学習（Speaker-Adaptive Training：SAT）法[9]を用いて，話者性を排除したバイリンガル話者空間の日本語，英語の音素セットを作成する方法を検討した．

4.1. 話者適応学習法

話者適応学習（以下，SAT）法は，不特定話者モデルにおける話者性の違いを排除したHMMのパラメータを計算する方法である．不特定話者モデルには作成に用いられる多数話者の話者性の情報が含まれており，そのために音素弁別機能が低下することが考えられる．SAT法ではモデルの学習時にMLLR話者適応を行い，それによって用いられる各話者の回帰行列を話者性とみなし，その回帰行列を用いて逆適応を行うことで基準話者モデルから話者性を排除している．その学習手順を図7に示す．

SAT法では，学習の最初の段階でMLLR話者適応を行う．この時得られた回帰行列を，その学習データの話者性とみなす．HMMの各パラメータの学習後，学習の始めに得られた回帰行列を用いて逆適応を行う．これにより，学習データにおける話者性を排除したモデル学習を行う．

4.2. 適応方法

SAT学習データにおけるある話者をrとする．学習の元データとなる音響モデル（種モデル）における，ある音素の平均ベクトルをμとおくと，話者rにおけるMLLR話者適応は，線形変換行列$A^{(r)}$と平行移動係数$b^{(r)}$を用いて，

$$\mu^{(r)'} = A^{(r)} \cdot \mu + b^{(r)}$$

このように表される．

この変換行列を用いて，平均ベクトルと対角共分散行列はSAT法により以下のように更新される．

$$\overline{\Sigma_{ik}} = \frac{\sum_{r,t} \gamma_{ik}^{(r)}(t) \cdot \left(o_t^{(r)} - \overline{\mu}_{ik}^{(r)}\right) \cdot \left(o_t^{(r)} - \overline{\mu}_{ik}^{(r)}\right)^T}{\sum_{r,t} \gamma_{ik}^{(r)}(t)}$$

$$\overline{\mu_{ik}} = \left\{ \sum_{r,t} \gamma_{ik}^{(r)}(t) \cdot A^{(r)T} \cdot \Sigma_{ik}^{(r)-1} \cdot A^{(r)} \right\}^{-1}$$

$$\times \left\{ \sum_{r,t} \gamma_{ik}^{(r)}(t) \cdot A^{(r)T} \cdot \Sigma_{ik}^{(r)-1} \cdot \left(o_t^{(r)} - \mu_{ik}^{(r)}\right) \right\}$$

γ_{ik}はHMMにおけるi状態k混合状態における入力学習データ$o_t^{(r)}$の出力確率，Σ_{ik}は種モデルの対角共分散行列を表す．平均ベクトルの更新式はMLLR法における回帰行列の導出式から導かれる．対角共分散行列の更新は，平均ベクトルの更新後に行われる．

このようにしてSAT法により学習された混合音響モデルは，バイリンガル話者全体を表す話者空間に写像されたものと考えられる．この音響モデルを，日本人学習話者の日本語発声に対する回帰行列を用いて話者適応を行う．

4.3. 性能評価実験

話者適応を行った音響モデルの評価として，2.3.節と同様の音素認識実験を行った．SAT法においては，2.3.節と同じバイリンガル話者3名の各1～50文章の発声を用いてMLLR回帰行列の導出，平均ベクトル，対角共分散行列の学習を行った．SAT法における繰り

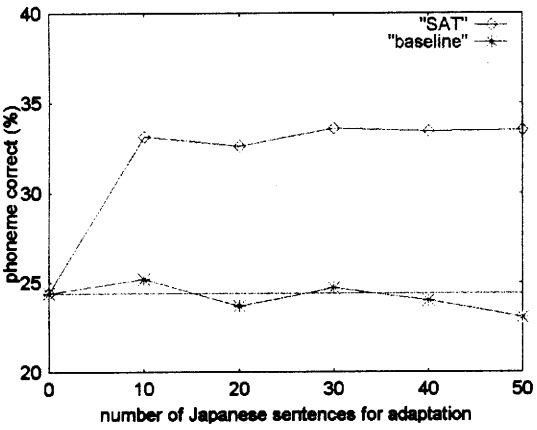

図８：音素認識結果（SAT法を用いた適応）

返し学習回数は1～20回とした．それ以外の条件は，すべて2.3.節に提示した条件と同じである．なお，学習の初期モデルとして，日本語英語混合不特定話者モデルを用いた．

バイリンガル話者各25文章の発声で7回繰り返し学習した音響モデルを適応したモデルの音素正解率を図8に示す．baselineは2.3.節におけるMLLR法による適応の音素正解率である．

バイリンガル話者の発声をSAT法で学習することにより，音素正解率は約10%上昇した．特定のバイリンガル話者空間を経由したものよりも音素正解率が高くなっている．SAT法を用いることで，話者性を排除した，バイリンガル話者全体の話者空間を作成することができたと考えられる．

SAT法において学習に用いる適応文章数は各話者20～30文章の時に音素正解率が高かった．また，繰り返し学習回数における音素正解率の有意な差は見られなかった．

5. まとめ

本報告では，外国語発音学習システムにおける認識率向上のための，二言語混合音響モデルの話者適応方法を検討した．日本人学習者が英語発声を学習することを想定し，日本語英語混合の音響モデルの話者適応を行い，その評価を行った．

混合音響モデルの話者適応において問題となるのは，言語によって不特定話者空間が異なることである．話者空間が異なったまま話者適応を行うと，認識精度が低下してしまう．そこで，不特定話者空間を同じにするために，バイリンガル話者の発声を用いた話者適応方法を提案した．複数のバイリンガル話者の発声からバイリンガル話者全体の話者空間を作成し，その空間を経由して話者適応を行うことで，音響モデルの精度が向上したことが確かめられた．

現状ではバイリンガル話者の発声を3名分しか入手できなかったので，今後はさらにデータ量を増やし，提案した適応方法のさらなる検証を行う必要がある．また，この適応方法を実装した英語発音学習システムを構築し，その認識精度を検証する予定である．

本研究では，日本語と英語の混合音響モデルに対して検討を行った．しかし，この方法を用いることで，日本語と英語以外の言語に対しても適応可能であり，多言語におけるCALLシステムに利用することができると考えられる．

文　献

[1] 竹林滋，渡邊末耶子，清水あつ子，斎藤弘子：“初級英語音声学”，大修館書店，1991

[2] 中村直生，中川聖一：“日本人の英語発音の評価法”，TECHNICAL REPORT OF IEICE（電子情報通信学会技報），SP2002-20（2002-05）pp13-18

[3] 大倉計美，杉山雅英，嵯峨山茂樹：“混合連続分布ＨＭＭを用いた移動ベクトル場平滑化話者適応方式”，信学技報，SP92-16 (1992)

[4] 伊藤彰則，長沢忠郎，鈴木基之，牧野正三：“日本語音声による話者適応を用いた英語韻律評価システム”，信学技報，SP2002-39（2002-06）pp19-24

[5] C. J.Leggetter，P. C. Woodland：“Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models”，Computer Speech and Language (1995) 9, pp171-185

[6] M. J. F. Gales, P. C. Woodland：“Mean and variance adaptation within the MLLR framework”，Computer Speech and Language (1996) 10，pp249-264

[7] Steev Young，et al.：“The HTK Book”，Cambridge University，1997

[8] 小笠原洋一，鈴木基之，伊藤彰則，牧野正三：“学習話者の異なる複数言語の音響モデルの話者適応の検討”，日本音響学会秋季研究会発表講演論文集（2003），3-6-4，pp.109-110

[9] T. Anastasakos，J，McDohough，R. Schwartz，J. Makhoul：“A Compact Model for Speaker-Adaptive Training”，Proc. ICSLP 96，vol.2，FrP2L1.3，1996